(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104615687 B
(45)授权公告日 2018.05.22

(21)申请号 201510033050.4

(22)申请日 2015.01.22

(65)同一申请的已公布的文献号
申请公布号 CN 104615687 A

(43)申请公布日 2015.05.13

(73)专利权人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路6号

(72)发明人 程学旗 王元卓 林海伦 贾岩涛 靳小龙 熊锦华 李曼玲 常雨骁 许洪波

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇 李科

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
US 8538916 B1,2013.09.17,
CN 103678316 A,2014.03.26,
CN 103034693 A,2013.04.10,
CN 104182463 A,2014.12.03,

审查员 窦广健

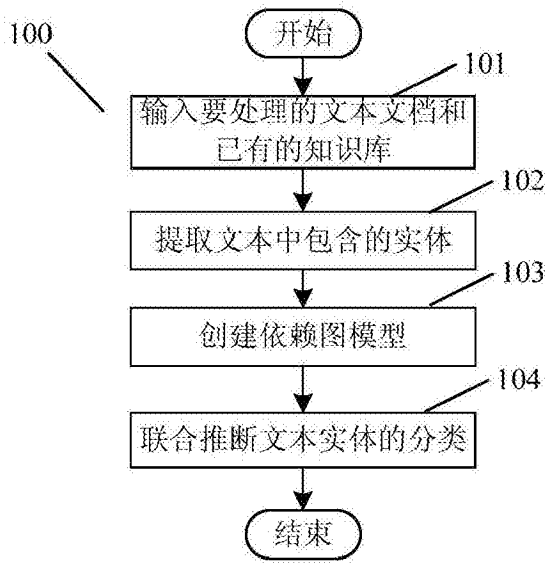
权利要求书3页 说明书9页 附图2页

(54)发明名称

一种面向知识库更新的实体细粒度分类方法与系统

(57)摘要

本发明提供一种面向知识库更新的实体细粒度分类方法与系统。所述方法包括：从文本中识别出实体；将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点构建依赖图，其中依赖图中的边的权值表示该边连接的两个结点之间的相关程度；以及，通过在所述依赖图上执行重启动随机游走，得到识别出的实体所属的分类。本发明能够克服现有技术在实体上下文缺乏的情况下难以实现对该实体进行细粒度分类的缺陷，并且提高了实体细粒度分类的准确率。



CN 104615687 B

1.一种面向知识库更新的实体细粒度分类方法,包括:

步骤1)、从文本中识别出实体;

步骤2)、将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点构建依赖图,其中依赖图中的边的权值表示该边连接的两个结点之间的相关程度;

所述步骤2)进一步包括:

步骤21)、根据语义相容度获得识别出的实体在知识库中的相关实体,并且获得该相关实体在知识库中的分类;其中,语义相容度表示识别出的实体的上下文信息与相关实体的描述文本的相似度;

步骤22)、将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点;

步骤23)、在表示识别出的实体的结点和表示相关实体的结点之间添加边,边的权值为该识别出的实体与该相关实体之间的语义相容度;

在表示相关实体的结点和表示分类的结点之间添加边,边的权值指示该相关实体是否属于该分类;

在表示相关实体的结点之间添加边,边的权值为该相关实体之间的语义相关度;

在表示分类的结点之间添加边,边的权值为该分类之间的相关度;

步骤3)、通过在所述依赖图上执行重启动随机游走,得到识别出的实体所属的分类。

2.根据权利要求1所述的方法,其中,根据下式计算语义相容度:

$$SC(em,e)=sim(X,T)=\frac{\vec{V}(X)\cdot\vec{V}(T)}{|\vec{V}(X)|\cdot|\vec{V}(T)|}$$

其中,SC(em,e)表示识别出的实体em与知识库中的相关实体e的语义相容度,X表示em的上下文信息,T表示e的描述文本,$\vec{V}(\cdot)$表示文本中包含的所有Biterm组成的TF-IDF向量,$|\vec{V}(\cdot)|$表示向量$\vec{V}(\cdot)$的模,Biterm表示文本中共现的词对。

3.根据权利要求2所述的方法,其中,识别出的实体的上下文信息是由出现在所述文本之前和之后的词组成的。

4.根据权利要求2或3所述的方法,其中,步骤21)包括:

将知识库中与识别出的实体的语义相容度大于0的实体作为相关实体。

5.根据权利要求1所述的方法,其中,根据下式计算相关实体之间的语义相关度:

$$SR(e_1,e_2)=1-\frac{\log(\max(|I_1|,|I_2|))-\log(|I_1\cap I_2|)}{\log(|Z|)-\log(\min(|I_1|,|I_2|))}$$

其中,SR(e1,e2)表示知识库中的相关实体e1和e2的语义相关度,I1和I2分别表示知识库中描述实体的文本中出现实体e1和e2的实体的集合,Z表示知识库中包含的所有实体的集合,|·|表示集合的大小。

6.根据权利要求1所述的方法,其中,根据下式计算分类之间的相关度:

$$CR(c_1,c_2)=\frac{|E_{c_1}\cap E_{c_2}|}{|E_{c_1}\cup E_{c_2}|}$$

其中，CR(c1,c2)表示分类c1和c2之间的相关度，Ec1和Ec2分别表示知识库中属于分类c1和c2的实体的集合，|·|表示集合的大小。

7.根据权利要求1-3中任何一个所述的方法，其中，步骤3)包括：

步骤31)、根据下式初始化所述依赖图中结点的分布状态：

$$\vec{r}_i^{(0)} = (r_{i(1)},...,r_{i(k)},...,r_{i(n)})$$

其中，n表示结点总数，$\vec{r}_i^{(0)}$表示结点i的初始分布状态；若k＝i，则$r_{i(k)}$＝1，否则$r_{i(k)}$＝0，k是自然数且1≤k≤n；

步骤32)、计算状态转移概率矩阵A＝(a_{ij})：

$$a_{ij} = \begin{cases} \dfrac{w_{ij}}{\sum_{k=1}^{n} u_{i(k)}} & \text{结点}i\text{和}j\text{邻接} \\ 0 & \text{结点}i\text{和}j\text{不邻接或}i=j \end{cases}$$

其中，a_{ij}表示在重启动随机游走过程中从结点i转移到结点j的概率，i、j是自然数且满足1≤i，j≤n；w_{ij}为结点i和结点j之间的边的权重；$\sum_{k=1}^{n} u_{i(k)}$表示连接结点i的所有边的权重之和；

步骤33)、对于每个结点，迭代地向其邻居结点进行状态转移，直到所述依赖图中每个结点的分布状态不随迭代次数的增加而改变；其中，在第t次迭代后结点i的分布状态$\vec{r}_i^{(t)}$表示如下：

$$\vec{r}_i^{(t)} = (1-\mu)A\vec{r}_i^{(t-1)} + \mu\vec{v}_i$$

其中，$\vec{r}_i^{(t)}$表示在第t次迭代后结点i的分布状态，t是自然数，i是自然数且1≤i≤n；$\vec{r}_i^{(t-1)}$表示在第t-1次迭代后结点i的分布状态；μ表示在第t次迭代后返回出发结点i的概率，μ为实数且0＜μ＜1；$\vec{v}_i$表示结点i的重启动向量且$\vec{v}_i = (v_{i(1)},...,v_{i(k)},...,v_{i(n)})$，若k＝i，则$v_{i(k)}$＝1，否则$v_{i(k)}$＝0，k是自然数且1≤k≤n；

步骤34)、根据结点的分布状态，得到其对应的分类。

8.根据权利要求7所述的方法，其中，步骤34)包括：

在表示识别出的实体的结点的分布状态中，将表示分类的结点按该结点对应的分量的值进行排序；

根据排序结果得到识别出的实体对应的分类。

9.一种面向知识库更新的实体细粒度分类系统，包括：

实体识别设备，用于从文本中识别出实体；

依赖图构建设备，用于将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点构建依赖图，其中依赖图中的边的权值表示该边连接的两个结点之间的相关程度；

同时，根据语义相容度获得识别出的实体在知识库中的相关实体，并且获得该相关实体在知识库中的分类；其中，语义相容度表示识别出的实体的上下文信息与相关实体的描述文本的相似度；

将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点；

并且，在表示识别出的实体的结点和表示相关实体的结点之间添加边，边的权值为该识别出的实体与该相关实体之间的语义相容度；

在表示相关实体的结点和表示分类的结点之间添加边，边的权值指示该相关实体是否属于该分类；

在表示相关实体的结点之间添加边，边的权值为该相关实体之间的语义相关度；

在表示分类的结点之间添加边，边的权值为该分类之间的相关度；以及

迭代设备，用于通过在所述依赖图上执行重启动随机游走，得到识别出的实体所属的分类。

一种面向知识库更新的实体细粒度分类方法与系统

技术领域

[0001] 本发明涉及信息处理技术领域，具体涉及一种面向知识库更新的实体细粒度分类方法与系统。

背景技术

[0002] 知识库是采用某种知识表示方式组织和管理的互相联系的知识集合。在知识工程领域，知识描述的要素一般包括分类、实体、关系、属性等要素，其中分类用于对知识库中的知识项进行语义分组或语义标注。知识库在很多领域中起到至关重要的作用，例如在信息检索中，知识库可以帮助搜索引擎理解用户查询、感知用户查询意图、进行查询扩展和查询问答等；此外，知识库在数据分析、舆情监控、深网资源发现等领域中也有广泛的应用。虽然目前存在众多知识库，但是它们在知识的覆盖率和时新性方面仍存在诸多限制，根本原因在于，随着大数据时代的到来，数据正以爆炸速度增长，在Web中每天都会产生新的知识。因此，为了构造高质量的知识库，将新产生的知识动态、实时、自动地更新到已有的知识库中，并且保障知识库的扩展能力、覆盖能力和时新性变得至关重要。

[0003] 实体作为知识描述的重要组成要素，知识库必然需要具备自动扩展实体的能力。要将新出现的实体更新到知识库中，需要首先确定实体在知识库中的位置，即实体在知识库中所属的分类信息。在确定了实体的分类之后，将新出现的实体添加到知识库的该分类下，从而丰富知识库中包含的实体集合。目前，实体分类方法主要有两类：实体粗粒度分类和实体细粒度分类。

[0004] 实体粗粒度分类将实体划分为粗粒度类别，如人名、地名、机构名等。主要采用有监督的方式训练实体分类模型，需要大量的人工标注的训练数据。这种方式无法直接应用到面向知识库的实体分类中，原因在于知识库将实体划分成成百上千个类别，它需要的训练数据的规模更大，而且创建如此规模的训练数据需要大量的人力。

[0005] 实体细粒度分类将实体划分为更细致的类别，主要采用启发式规则或基于弱监督的方法对实体进行分类。其中，基于启发式规则的方法直接通过定义的句法模式为实体进行类别标注，这种方法操作简单，但是需要人工维护和定义大量的规则。基于弱监督的方法提取实体的上下文，利用上下文的词法、句法特征计算实体所属的分类信息，然而这种方法的准确率较低，而且这种方法在上下文缺乏的情况下将难以推断实体的分类信息。

[0006] 综上所述，现有的实体粗粒度分类方法并不适用于知识库的更新，而现有的实体细粒度分类方法准确率较低。

发明内容

[0007] 为解决上述问题，根据本发明的一个实施例，提供一种面向知识库更新的实体细粒度分类方法，包括：

[0008] 步骤1)、从文本中识别出实体；

[0009] 步骤2)、将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的

分类作为结点构建依赖图，其中依赖图中的边的权值表示该边连接的两个结点之间的相关程度；

[0010] 步骤3)、通过在所述依赖图上执行重启动随机游走，得到识别出的实体所属的分类。

[0011] 上述方法中，步骤2)包括：

[0012] 步骤21)、根据语义相容度获得识别出的实体在知识库中的相关实体，并且获得该相关实体在知识库中的分类；其中，语义相容度表示识别出的实体的上下文信息与相关实体的描述文本的相似度；

[0013] 步骤22)、将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点；

[0014] 步骤23)、在表示识别出的实体的结点和表示相关实体的结点之间添加边，边的权值为该识别出的实体与该相关实体之间的语义相容度；

[0015] 在表示相关实体的结点和表示分类的结点之间添加边，边的权值指示该相关实体是否属于该分类；

[0016] 在表示相关实体的结点之间添加边，边的权值为该相关实体之间的语义相关度；

[0017] 在表示分类的结点之间添加边，边的权值为该分类之间的相关度。

[0018] 上述方法中，根据下式计算语义相容度：

[0019] $$SC(em,e)=sim(X,T)=\frac{\vec{V}(X)\cdot\vec{V}(T)}{|\vec{V}(X)|\cdot|\vec{V}(T)|}$$

[0020] 其中，SC(em,e)表示识别出的实体em与知识库中的相关实体e的语义相容度，X表示em的上下文信息，T表示e的描述文本，$\vec{V}(\cdot)$表示文本中包含的所有Biterm组成的TF-IDF向量，$|\vec{V}(\cdot)|$表示向量$\vec{V}(\cdot)$的模，Biterm表示文本中共现的词对。其中，识别出的实体的上下文信息是由出现在所述文本之前和之后的词组成的。

[0021] 上述方法中，步骤21)包括：

[0022] 将知识库中与识别出的实体的语义相容度大于0的实体作为相关实体。

[0023] 上述方法中，根据下式计算相关实体之间的语义相关度：

[0024] $$SR(e_1,e_2)=1-\frac{\log(\max(|I_1|,|I_2|))-\log(|I_1\cap I_2|)}{\log(|Z|)-\log(\min(|I_1|,|I_2|))}$$

[0025] 其中，SR(e_1,e_2)表示知识库中的相关实体e_1和e_2的语义相关度，I_1和I_2分别表示知识库中描述实体的文本中出现实体e_1和e_2的实体的集合，Z表示知识库中包含的所有实体的集合，|·|表示集合的大小。

[0026] 上述方法中，根据下式计算分类之间的相关度：

[0027] $$CR(c_1,c_2)=\frac{|E_{c_1}\cap E_{c_2}|}{|E_{c_1}\cup E_{c_2}|}$$

[0028] 其中，CR(c_1,c_2)表示分类c_1和c_2之间的相关度，E_{c_1}和E_{c_2}分别表示知识库中属于

分类c_1和c_2的实体的集合，|·|表示集合的大小。

[0029] 上述方法中，步骤3)包括：

[0030] 步骤31)、根据下式初始化所述依赖图中结点的分布状态：

[0031] $$\vec{r}_i^{(0)} = (r_{i(1)},\ldots,r_{i(k)},\ldots,r_{i(n)})$$

[0032] 其中，n表示结点总数，$\vec{r}_i^{(0)}$表示结点i的初始分布状态；若k＝i，则$r_{i(k)}$＝1，否则$r_{i(k)}$＝0，k是自然数且1≤k≤n；

[0033] 步骤32)、计算状态转移概率矩阵A＝(a_{ij})：

[0034] $$a_{ij} = \begin{cases} \dfrac{w_{ij}}{\sum_{k=1}^{n} u_{i(k)}} & \text{结点}i\text{和}j\text{邻接} \\ 0 & \text{结点}i\text{和}j\text{不邻接或}i=j \end{cases}$$

[0035] 其中，a_{ij}表示在重启动随机游走过程中从结点i转移到结点j的概率，i、j是自然数且满足1≤i，j≤n；w_{ij}为结点i和结点j之间的边的权重；$\sum_{k=1}^{n} u_{i(k)}$表示连接结点i的所有边的权重之和；

[0036] 步骤33)、对于每个结点，迭代地向其邻居结点进行状态转移，直到所述依赖图中每个结点的分布状态不随迭代次数的增加而改变；其中，在第t次迭代后结点i的分布状态$\vec{r}_i^{(t)}$表示如下：

[0037] $$\vec{r}_i^{(t)} = (1-\mu)A\vec{r}_i^{(t-1)} + \mu\vec{v}_i$$

[0038] 其中，$\vec{r}_i^{(t)}$表示在第t次迭代后结点i的分布状态，t是自然数，i是自然数且1≤i≤n；$\vec{r}_i^{(t-1)}$表示在第t-1次迭代后结点i的分布状态；μ表示在第t次迭代后返回出发结点i的概率，称为重启动因子，μ为实数且$0<\mu<1$ $\vec{v}_i$表示结点i的重启动向量且$\vec{v}_i = (v_{i(1)},\ldots,v_{i(k)},\ldots,v_{i(n)})$，若k＝i，则$v_{i(k)}$＝1，否则$v_{i(k)}$＝0，k是自然数且1≤k≤n；

[0039] 步骤34)、根据结点的分布状态，得到其对应的分类。

[0040] 上述方法中，步骤34)包括：

[0041] 在表示识别出的实体的结点的分布状态中，将表示分类的结点按该结点对应的分量的值进行排序；

[0042] 根据排序结果得到识别出的实体对应的分类。

[0043] 根据本发明的一个实施例，还提供一种面向知识库更新的实体细粒度分类系统，包括：

[0044] 实体识别设备，用于从文本中识别出实体；

[0045] 依赖图构建设备，用于将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点构建依赖图，其中依赖图中的边的权值表示该边连接的两个结点之间的相关程度；以及迭代设备，用于通过在所述依赖图上执行重启动随机游走，得到识别出的实体所属的分类。

[0046] 本发明能够克服现有技术在实体上下文缺乏的情况下难以实现对该实体进行细粒度分类的缺陷，通过建模同一文本中出现的实体之间的语义相关性，以及文本实体与知识库实体及其分类之间的关系，利用该语义相关性和关系为同一文本中实体细粒度分类提供有力的证据支持，并且通过重启动随机游走算法，提升了实体细粒度分类的准确率。

附图说明

[0047] 以下参照附图对本发明实施例作进一步说明，其中：

[0048] 图1是根据本发明一个实施例的面向知识库更新的实体细粒度分类方法的流程图；

[0049] 图2是根据本发明一个实施例的创建依赖图模型的方法的流程图；

[0050] 图3是根据本发明一个实施例的依赖图的示例；

[0051] 图4是根据本发明一个实施例的联合推断实体分类的方法的流程图。

具体实施方式

[0052] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图通过具体实施例对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0053] 根据本发明的一个实施例，提供一种面向知识库更新的实体细粒度分类方法。

[0054] 概括而言，该方法包括：从文本中识别出实体；将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点构建依赖图，其中依赖图中的边的权值表示该边连接的两个结点之间的相关程度；以及，通过在所述依赖图上执行重启动随机游走，得到识别出的实体所属的分类。该方法基于分布假设理论，即两个实体所在的上下文语义相关性越大，则它们属于同一类别的可能性越大。

[0055] 现参考图1，描述本发明方法的各个步骤。

[0056] 步骤101：输入要进行处理的文本文档和目标知识库

[0057] 选择要进行处理的文本文档D和目标知识库KB，初始化系统输入。

[0058] 如上文所述，知识库(Knowledge Base，KB)是由描述知识的实体、分类、关系、属性等要素组成的，因此可将目标知识库KB建模成如下形式：

[0059] KB＝<C,E,P,R>

[0060] 其中，C表示目标知识库中包含的分类集合；E和P分别表示属于分类的实体及其属性的集合，R是定义分类、实例、属性之间的关系的函数。在集合E中，每一个实体e可用如下形式来表示：

[0061] e＝<name,aliases,T>

[0062] 其中，name表示实体e的名字；aliases表示实体e的别名的集合；T表示实体e的描述文本。实体e的属性集合P_e和实体e所属的分类集合C_e可通过知识库KB的函数R求得，且满足$P_e \subseteq P, C_e \subseteq C$。

[0063] 可利用现有的各种百科数据库资源来建模上述形式的目标知识库，例如，在本步骤中采用基于维基百科创建的知识库作为输入的目标知识库。

[0064] 步骤102：提取文本文档中包含的实体

[0065] 利用命名实体识别工具，提取文本文档D中包含的所有实体的集合。

[0066] 文本文档D中包含的所有实体的集合可记为：

[0067] $EM=\{em_i|i为整数,0\leq i\leq|D|\}$

[0068] 其中，|D|为文本文档的长度；集合中的每一个元素em用如下形式来表示：

[0069] $em=\langle name,D,X\rangle$

[0070] 其中，name表示em的名字；D表示em的来源文本文档，X表示描述em的上下文。在一个实施例中，用em出现在文本文档D周围的词窗口来表示X，窗口大小为k(k为整数且$0<k\leq|D|$)，即上下文X的长度为2k(X是由出现在文本文档D前的k个词和文本文档D后的k个词组成的)，优选地，$k=\min(50,|D|)$。

[0071] 本领域技术人员应理解，可利用现有的各种命名实体识别工具来提取文本中的实体。在一个实施例中，利用Stanford NER作为命名实体识别工具。

[0072] 步骤103：创建依赖图

[0073] 根据从文本文档D中提取的实体集合EM和目标知识库KB，创建依赖图，从而统一建模文本文档D中不同实体之间的语义相关性，以及文本文档D中的实体与知识库KB中的实体及其所属分类之间的依赖关系。

[0074] 参考图2，在一个实施例中，创建依赖图包括以下子步骤：

[0075] 步骤1031：输入从文本文档D中识别的实体集合EM和目标知识库KB。

[0076] 步骤1032：选择候选实体。

[0077] 根据描述实体的文本的语义相容度(Semantic Compatibility，SC)，在知识库KB中为每一个实体$em\in EM$选择与其语义相容的候选实体集合，记为：

[0078] $ES_{em}=\{e\in E|SC(em,e)>0\}$

[0079] 其中，SC(em,e)表示em和知识库实体e之间的语义相容度。在一个实施例中，采用基于Biterm的余弦相似度的方式计算该语义相容度：

[0080] $$SC(em,e)=sim(X,T)=\frac{\vec{V}(X)\cdot\vec{V}(T)}{|\vec{V}(X)|\cdot|\vec{V}(T)|}$$

[0081] 其中，SC(em,e)为实数且$0\leq SC(em,e)\leq 10$；X为描述em的上下文信息；T为e的描述文本；sim(X,T)为X与T的相似度；$\vec{V}(\cdot)$为文本中包含的所有的Biterm组成的TF-IDF向量，$|\vec{V}(\cdot)|$为向量的模，且一个Biterm为文本中的一个共现的词对。例如，给定文本“苹果应用商店”，该文本通过分词获得三个词“苹果”、“应用”、“商店”，那么该文本包含的Biterm集合为{苹果应用，苹果商店，应用商店}。

[0082] 根据上述公式，若$SC(em,e)>0$，则选择e作为em的候选实体，从而获得与em语义相容的候选实体集合ES_{em}。

[0083] 步骤1033：选择候选分类。

[0084] 根据知识库KB中的关系定义函数R，获得步骤1032中选择的每个候选实体e在知识库KB中所属的分类的集合$C_e\subseteq C$，将其作为候选分类集合。

[0085] 步骤1034：确立依赖图中的结点及连边信息。

[0086] 依赖图中的结点集合包括从文本文档D中提取的所有实体(简称文本实体)的集

合、与提取的实体语义相容的候选实体(简称知识库实体)的集合,以及候选实体所属的分类(简称知识库分类)的集合。

[0087] 在确立图中的结点之后,在这些结点之间分配连边及权重,具体包括:

[0088] 1、在代表文本实体em的结点和代表与其语义相容的知识库实体结点e之间添加连边,边上的权值为它们之间的语义相容度SC(em,e)。

[0089] 2、在代表知识库实体e的结点和代表其所属的分类c的结点之间添加连边,边上的权值为它们之间的所属关系(Attachment Relatedness,AR),若实体属于该分类,则权值为1.0,若不属于则权值为0.0。

[0090] 3、在代表知识库实体的两个结点e_1和e_2之间添加连边,边上的权值为它们之间的语义相关度(Semantic Relatedness,SR)。值得注意的是,在此通过知识库实体之间的语义相关性来间接度量同一文本中实体之间的语义相关性。

[0091] 在一个实施例中,基于规范化的谷歌距离(google distance)计算实体e_1和e_2之间的语义相关度SR(e_1,e_2):

[0092] $$SR(e_1,e_2)=1-\frac{\log(\max(|I_1|,|I_2|))-\log(|I_1\cap I_2|)}{\log(|Z|)-\log(\min(|I_1|,|I_2|))}$$

[0093] 其中,SR(e_1,e_2)为实数且0≤SR(e_1,e_2)≤1.0;I_1和I_2分别表示知识库KB中,描述实体的文本中出现实体e_1和e_2的实体的集合,Z表示知识库KB中包含的所有实体的集合,|·|表示集合的大小。

[0094] 4、在代表知识库分类的两个结点c_1和c_2之间添加连边,边上的权值为它们之间的相关程度(Correlation,CR)。在一个实施例中,采用Jaccard系数计算分类c_1和c_2之间的相关度CR(c_1,c_2):

[0095] $$CR(c_1,c_2)=\frac{|E_{c_1}\cap E_{c_2}|}{|E_{c_1}\cup E_{c_2}|}$$

[0096] 其中,CR(c_1,c_2)为实数且0≤CR(c_1,c_2)≤1.0,E_{c_1}和E_{c_2}分别表示知识库KB中属于分类c_1和c_2的实体的集合,|·|表示集合的大小。

[0097] 通过确立结点和连边,构造了关于文本文档D中所有实体EM的依赖图,记为G=(V,E,W)。G是一个无向图,其中V为图的顶点集合,包括给定文本中所有的实体、与这些实体语义相容的知识库中的所有实体,以及这些实体所属分类的集合。E为这些结点之间的边集合;W:E→R(R是实数)为边上的权值。

[0098] 给定一段文本“对于球员来说,名人堂是伟大的丰碑,也是对于球员生涯的肯定,是除了冠军戒指之外最好的认可。但是因为球员想进入名人堂都必须等到退役后5年,所以飞人直到2009年才等到了这样的殊荣。不过,这并不妨碍乔丹的名字在NBA乃至全世界篮坛的历史上闪闪发光”。利用命名实体识别工具识别出3个不同的实体:“名人堂”、“乔丹”、“NBA”。利用本发明提供的方法,对这3个实体创建依赖图模型。如图3所示,图中总共包含12个结点:3个文本实体,6个知识库实体和3个知识库分类,并且包含12条边。

[0099] 步骤104:根据创建的依赖图,联合推断实体的分类信息

[0100] 在上一步创建的依赖图上,执行随机游走算法,如重启动随机游走算法。不断迭代

地在依赖图上做随机游走，直至图中结点的分布状态不随着迭代次数的增加而改变，即达到稳定状态为止。此时，根据代表文本实体的结点的分布状态，获得其对应的分类标签，从而推断出文本实体的细粒度分类信息。

[0101] 下面将参考图4，结合本发明的一个实施例，对本步骤进行具体说明：

[0102] 步骤1041：初始化算法输入。

[0103] 输入创建的依赖图G＝(V,E,W)。

[0104] 步骤1042：初始化依赖图中的结点的分布状态。

[0105] 记图G中结点的数目为n＝|V|，边的数目为m＝|E|，G中的结点编号分别为1，…，i，…，n(i为自然数且1≤i≤n)。

[0106] 设置算法初始时依赖图中结点i的分布状态$\vec{r}_i^{(0)}$，该分布状态是关于图G中包含的所有结点的一个n x 1维的列向量，其中n是图G中结点的数目。$\vec{r}_i^{(0)}$记为：

[0107] $$\vec{r}_i^{(0)}=(r_{i(1)},\ldots,r_{i(k)},\ldots,r_{i(n)})$$

[0108] 其中，对于该向量中的每一个分量$r_{i(k)}$的取值如下：若k＝i，则$r_{i(k)}$＝1，否则$r_{i(k)}$＝0，k是自然数且1≤k≤n。

[0109] 步骤1043：根据依赖图G＝(V,E,W)的邻接矩阵U＝(u_{ij})，计算随机游走过程中状态转移概率矩阵A＝(a_{ij})，i、j是自然数且满足1≤i，j≤n。对于邻接矩阵U，u_{ij}取值如下：

[0110] $$u_{ij}=\begin{cases}w_{ij} & \text{结点}i\text{和}j\text{邻接}\\ 0 & \text{结点}i\text{和}j\text{不邻接或}i=j\end{cases}$$

[0111] 其中，w_{ij}为结点i和结点j之间的连边上的权重，由G＝(V,E,W)中的W:E→R(R是实数)来确定。

[0112] 对于状态转移概率矩阵A，a_{ij}表示在重启动随机游走过程中，从结点i转移到结点j的概率。记图G＝(V,E,W)中结点i与其他所有结点组成的邻接向量为$\vec{u}_i=(u_{i(1)},\ldots,u_{i(k)},\ldots,u_{i(n)})$，邻接向量即为邻接矩阵U中第i行元素组成的向量，k是自然数且1≤k≤n。根据结点i的邻接向量，按照如下方式计算a_{ij}：

[0113] $$a_{ij}=\begin{cases}\dfrac{w_{ij}}{\sum_{k=1}^{n}u_{i(k)}} & \text{结点}i\text{和}j\text{邻接}\\ 0 & \text{结点}i\text{和}j\text{不邻接或}i=j\end{cases}$$

[0114] 从上式可知，若i＝j或者结点i和j之间不存在连边，则a_{ij}＝0；若结点i和j之间存在连边，则a_{ij}的取值与结点i和结点j之间边上的权重成比例，即为结点i和结点j之间连边上的权重与连接结点i的所有连边上的权重和之比。

[0115] 步骤1044：在依赖图上，从出发结点i开始，不断迭代地向其周围的邻居结点进行状态转移。在第t次迭代后，图中结点i的分布状态$\vec{r}_i^{(t)}$表示如下：

[0116] $$\vec{r}_i^{(t)}=(1-\mu)A\vec{r}_i^{(t-1)}+\mu\vec{v}_i$$

[0117] 其中，t为自然数；$\vec{r}_i^{(t-1)}$表示在第t-1次迭代后结点i的分布状态；$\vec{r}_i^{(t)}$表示在第t次迭代后结点i的分布状态；μ表示在第t次迭代后返回出发结点i的概率(称为重启动因子，μ为实数且0<μ<1，优选为0.15)；$\vec{v}_i$是结点i的重启动向量，是一个关于图G中包含的所有结点的一个n x 1维的列向量，n是图G中结点的数目，$\vec{v}_i$记为$\vec{v}_i=(v_{i(1)},...,v_{i(k)},...,v_{i(n)})$，其中向量中的每一个分量$v_{i(k)}$的取值如下：若k＝i，则$v_{i(k)}$＝1，否则$v_{i(k)}$＝0，k是自然数且1≤k≤n。

[0118] 重复执行步骤1044，直至依赖图中每个结点i(i为自然数且1≤i≤n)的分布状态$\vec{r}_i$达到稳定时终止算法。也就是说，依赖图中的结点i的分布状态$\vec{r}_i$不再随着迭代次数t的增加而改变(结点的分布达到稳定状态)。此时，根据代表文本实体的结点的分布状态，获得其对应的分类标签，从而推断出该文本实体具体的分类信息。

[0119] 具体地，根据上文所讨论的，$\vec{r}_i^{(0)}$是关于图G中包含的所有结点的一个n x 1维的列向量。对于达到稳定状态的结点i的分布状态$\vec{r}_i^{(t)}$，其也是关于图G中包含的所有结点的一个n x 1维的列向量，因此图G中的分类结点也包含在这个向量中。在向量$\vec{r}_i^{(t)}$中，分类结点所对应的分量的值作为通过重启动随机游走之后结点i表示的实体属于该分类的概率值，通过概率排序可获得结点i表示的实体所对应的分类标签(即选择最大概率对应的分类)。

[0120] 联合推断文本实体的分类是为了利用知识库的分类信息对文本实体标注其所属的知识库分类，通过同一文本中一个实体分类的推断对另一个实体的分类推断的相互促进作用，同时实现对同一文本中所有实体的分类的推断。

[0121] 根据本发明的一个实施例，还提供一种面向知识库更新的实体细粒度分类系统，包括实体识别设备、依赖图构建设备和迭代设备。

[0122] 其中实体识别设备用于从文本中识别出实体，例如，如上文所述的命名实体识别工具。依赖图构建设备用于将识别出的实体、知识库中与其相关的实体以及相关实体在知识库中的分类作为结点构建依赖图。迭代设备用于通过在所述依赖图上执行重启动随机游走，得到识别出的实体所属的分类。

[0123] 为验证本发明提供的面向知识库更新的实体细粒度分类方法与系统的有效性，发明人分别采用现有最新的实体分类技术(APOLLO)和本发明提供的方法，在真实YAGO数据集上进行了实验，实验参数如下：

[0124] 实验所用的实体是利用YAGO中person分类的15个子目录下随机选择出来的数据组成的，其中从每个目录下随机最多选择200个实体，共计选择出2650个实体作为最终的数据集DSec。设置DSec中用作训练的数据占总数据的比例ρ＝0.8，迭代次数t＝10，重启动因子μ＝0.15，窗口大小k＝50。

[0125] 经过实验得到如下结果：采用现有APOLLO技术的分类准确率为0.7254，而采用本发明提供的方法和系统所得到分类结果的准确率为0.7708。采用本发明提供的实体细粒度分类方法和系统与采用现有的APOLLO技术相比，准确率提升了4.5％左右。

[0126] 综上，本发明提供了一种面向知识库更新的实体细粒度分类方法与系统，该方法基于依赖图，建模同一文本中出现的实体之间的语义相关性，并利用此相关性为同一文本

中实体细粒度的分类提供有力的证据支持，通过基于重启动随机游走算法的联合推断方法，实现实体细粒度分类的准确率的提升。

[0127] 应当理解，虽然本说明书是按照各个实施例描述的，但并非每个实施例仅包含一个独立的技术方案，说明书的这种叙述方式仅仅是为清楚起见，本领域技术人员应当将说明书作为一个整体，各实施例中的技术方案也可以经适当组合，形成本领域技术人员可以理解的其他实施方式。

[0128] 以上所述仅为本发明示意性的具体实施方式，并非用以限定本发明的范围。任何本领域的技术人员，在不脱离本发明的构思和原则的前提下所作的等同变化、修改与结合，均应属于本发明保护的范围。

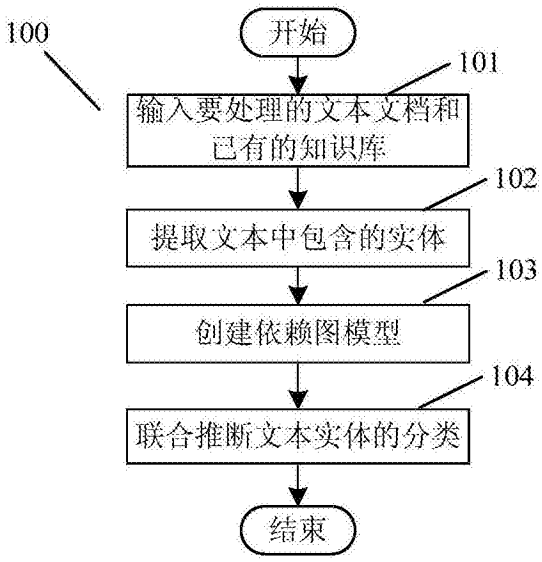

100
开始
101
输入要处理的文本文档和已有的知识库
102
提取文本中包含的实体
103
创建依赖图模型
104
联合推断文本实体的分类
结束

图1

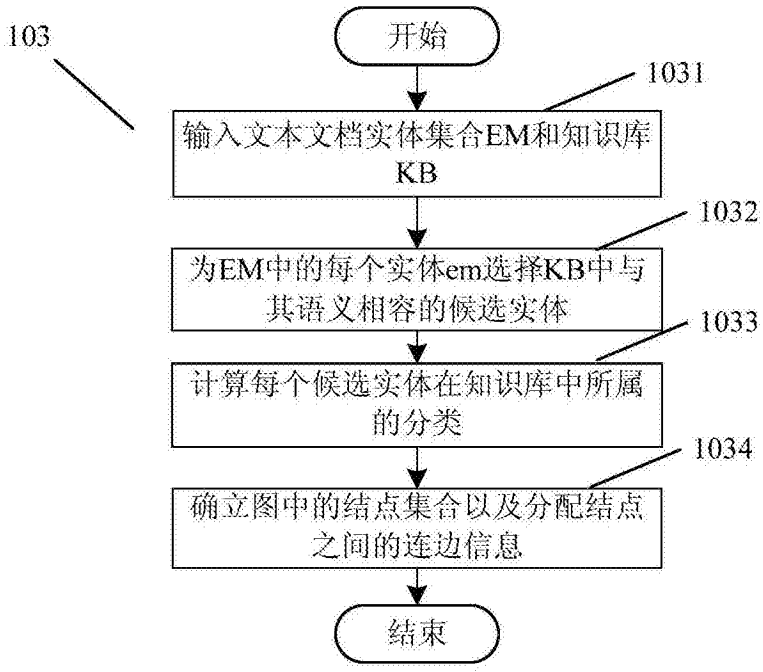

103
开始
1031
输入文本文档实体集合EM和知识库KB
1032
为EM中的每个实体em选择KB中与其语义相容的候选实体
1033
计算每个候选实体在知识库中所属的分类
1034
确立图中的结点集合以及分配结点之间的连边信息
结束

图2

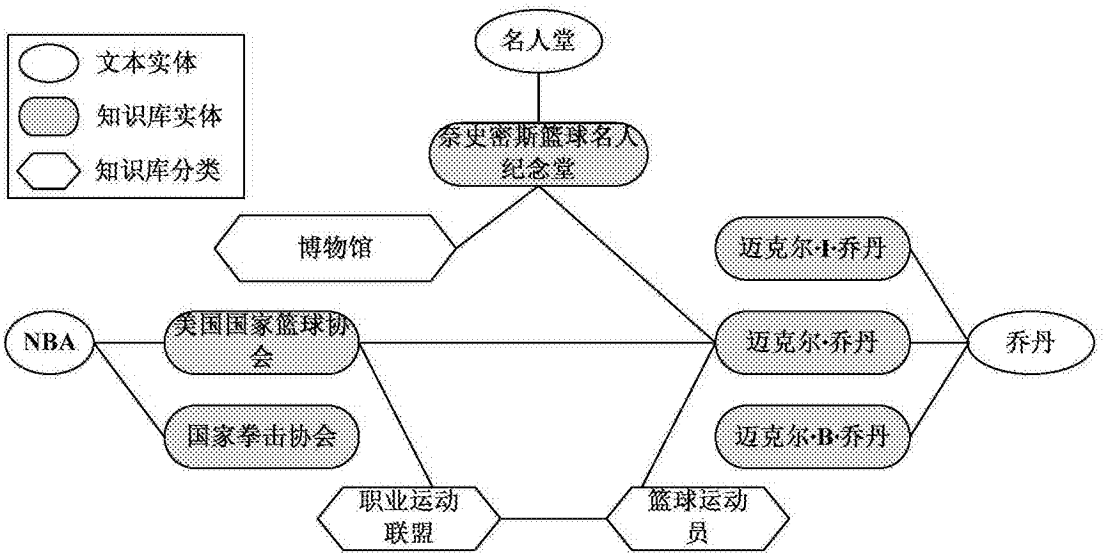

文本实体
知识库实体
知识库分类
名人堂
奈史密斯篮球名人纪念堂
博物馆
迈克尔·I·乔丹
NBA
美国国家篮球协会
迈克尔·乔丹
乔丹
国家拳击协会
迈克尔·B·乔丹
职业运动联盟
篮球运动员

图3

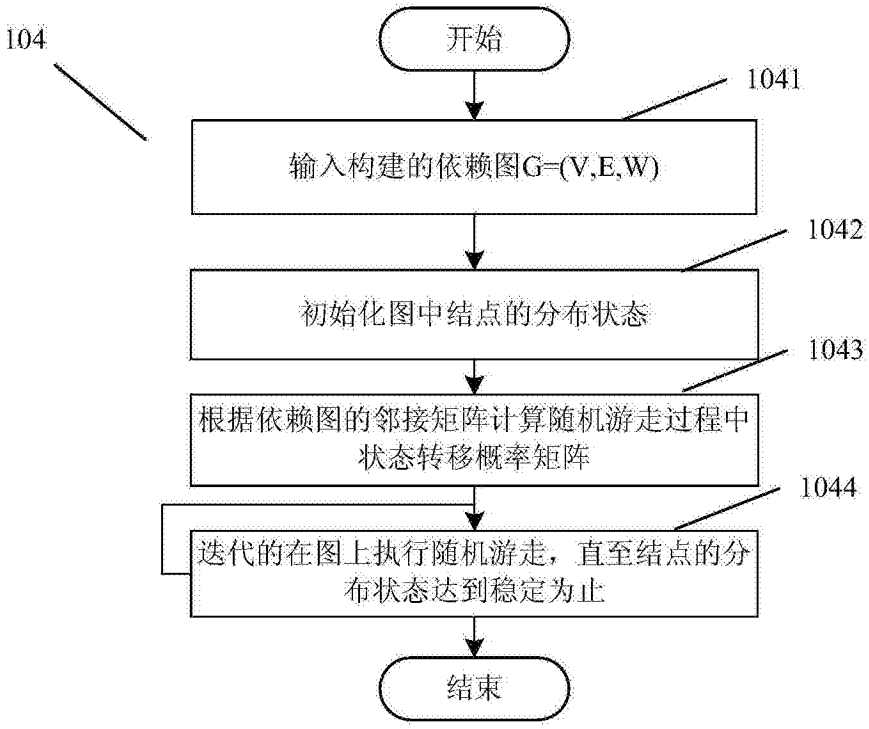

104
开始
1041
输入构建的依赖图G=(V,E,W)
1042
初始化图中结点的分布状态
1043
根据依赖图的邻接矩阵计算随机游走过程中状态转移概率矩阵
1044
迭代的在图上执行随机游走，直至结点的分布状态达到稳定为止
结束

图4